Calibration of Speech Data Acquisition Path

Field of Invention

This invention relates to speech recognition and more particularly to calibration of speech data acquisition path.

Background of Invention

A speech acquisition path refers to the whole speech transmission path before the speech is actually digitized

Typical speech acquisition path includes therefore air from lips to the microphone, microphone, wires, antialiasing filters, analog-to-digital converter. This is determining the transfer function of the system. Noises can be introduced at each of these devices and from power supply of the analog-to-digital converter.

Practical speech acquisition lines, especially for low cost devices, introduce both convolutive and additive noises to the input speech, and cause additional statistical mismatch between an utterance to be recognized and trained speech model set. Such mismatch will cause performance degradation.

Previously, SNR-dependent cepstral normalization (SDCN), fixed-code-word-dependent cepstral normalization (FCDCN) [See A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Kluwer Academic Publishers, 1993], multi-variate Gaussian based cepstral normalization [P. Moreno, B. Raj, and R. Stern. Multi-variate Gaussian based cepstral normalization. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing,* Detroit, 1995] and statistical re-estimation [P. Moreno, B. Raj, and R. Stern. A unified approach to robust speech recognition. In *Proceedings of European Conference on Speech Communication and Technology,* Madrid, Spain, Sept. 1995] have been proposed to deal with similar problem. They all assume that the distortions can be modeled by a bias in the cepstral domain, which is clearly not the case for additive distortions. Vector Taylor series has been used to approximate the distortion as function of cepstral representation of additive and convolutive noises. See reference P. J. Moreno, B. Raj, and R. M. Stern. A vector taylor series

approach for environment-independent speech recognition. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing,* Atlanta, GA, 1996.

Summary of Invention

In accordance with one embodiment of the present invention the parameters of the convolutive and additive noise are determined by a calibration of the speech acquisition path to compensate for the mismatch between an utterance to be recognized and a trained speech model set. In the power spectrum domain both types of noises are modeled as polynomial functions of frequency. The model parameters are estimated with maximum likelihood (ML) criterion on a set of simultaneous recordings.

Description of Drawings

Fig. 1 illustrates the equipment configuration for the calibration according to one embodiment of the present invention.

Fig. 2 illustrates the process steps for the calibration;

Fig. 3 illustrates the results of estimation for convolutive noise estimation at 30db SNR;

Fig. 4 .illustrates the results of estimation for additive noise elimination at 30db SNR;

Fig.5 illustrates the results of estimation for convolutive noise estimation at 24db SNR;

Fig.6 illustrates the results of estimation for additive noise elimination at 24db SNR;

Fig. 7 illustrates that the estimation of the convolutive bias with independent component model gives 4.7 to 8.4 time larger estimation error.

Fig. 8 illustrates that the estimation of additive bias with independent component model gives 7.5 to 12.3 time larger estimation error than with polynomial models.

Description of Preferred Embodiments

In the power-spectral domain, the additive and convolutive noises are modeled as polynomial functions of frequency. The model parameters are estimated with Maximum Likelihood (ML) criterion, on a set of simultaneous recordings.

Once the parameter of the convolutive and additive noises are determined, speech recognizers can be compensated with these parameters, either by speech enhancement, or by model adaptation. See references: M. J. F. Gales. "nice" model-based compensation schemes for robust speech recognition. In *Robust speech recognition for unknown communication channels,* pages 55-64, Pont-a-mousson, France, 1997; Y Gong. Speech recognition in noisy environments: A survey. *Speech Communication,* 16(3):261-291, April 1995; and C.H. Lee. On feature and model compensation approach to robust speech recognition. In *Robust speech recognition for unknown communication channels,* pages 45-54, Pont-a-mousson, France, 1997.

The present invention is determining the transfer function and associated noise estimate. What we want to determine is the frequency response of the microphone as well as the noise that may have been introduced by the A/D (analog to digital)conversion process. This is particularly needed for low quality microphones and noisy systems. To model the transfer function we need to determine the microphone followed by H (the linear filter) and the noise N of the A/D system. Hardware illustrated in Fig. 1 is used.

The equipment 10 used is outlined in Fig. 1. It produces two signals: reference Y_R and noisy Y_N. Y_R is assumed to give a digitized speech signal under the same acoustic environment of the speech used to train the recognizer and is represented by microphone 11 and DAT (high quality data recording) equipment 15. For this reference H is 1 and N is zero. Y_N is the noisy speech signal under the acoustic environment of the target application with the test microphone 13 and equipment under test path 17. Reference microphone 11 is the one used for recording the training database, and test microphone 13 is the one to be used in the target product for field operation. The equipment test path 17 may introduce both convolutive noises (microphone, pre-A/D (pre analog to digital) filter)and additive noises (environmental noises, any noises introduced before A/D, etc.), even when there is no background noise. The purpose here is to estimate the two noise components, with very short speech utterance (typically one) represented by X.

In the equipment configuration Fig.1, we identify two environments: reference (R) and noisy (N). We represent a signal by its power spectral density sampled with a DFT (Discrete Fourier Transform) of *2M d*imensions. Using Y_R and Y_N the distortion parameters are determined. For a given signal X, reference signal is

$$Y_R \triangleq \left[Y_R^{(1)}, Y_R^{(2)}, \ldots Y_R^{(M)}\right]^t = H_R X \tag{1}$$

and that of the noisy signal is:

$$Y_N \triangleq \left[Y_N^{(1)}, Y_N^{(2)}, \ldots Y_N^{(M)}\right]^t = H_N X + N_N + \mathrm{e} \tag{2}$$

where e is assumed to be a zero mean Gaussian distribution with diagonal covariance matrix, i.e., e : [0, *RJ*.

As the pre-A/D filtering 15a as well as other parts of the reference equipment are of much higher acoustic quality than that of most speech recognizers, it is assumed that H_R contains the information on the reference microphone 11. H_N models the test microphone 13 and the pre A/D filter 17a of the equipment test path N_N models the noise background introduced at any point of the test equipment. H_R and H_N are both M-dimensional diagonal matrices, and N_N a M-dimensional vector. One can remove X by substituting $H_R^{-1} Y_R$ for X.

From Equation 1 and Equation 2, we have:

$$Y_N = H_\Delta Y_R + N_N + \mathrm{e} \tag{3}$$

with

$$H_{\Delta} \underline{\Delta} H_N H_R^{-1} \text{(ratio of transfer function)} \tag{4}$$

We are interested in the estimate acoustic changes represented by H_Δ and N_N. There are so many combinations to determine these values. A model is assumed herein that Y_N is only true when Y_N follows a Gaussian distribution as in equation 5.

Let the i-th observed Y_N be $Y_N(i)$, and the i-th observed Y_R be $Y_R(i)$. The likelihood of a set of T observed noisy signals is then:

$$P(Y_N(1), Y_N(2), \ldots Y_N(T) \mid \lambda) = \prod_{i=1}^{T} \frac{1}{(2\pi)^{M/2} \mid R \mid^{1/2}} \exp^{-\frac{1}{2}(Y_N(i)-H_\Delta Y_R(i)-N_N)' R^{-1} (Y_N(i)-H_\Delta Y_R(i)-N_N)} \tag{5}$$

With limited amount of data, direct estimation of the parameters $H_\Delta \in R_M$ and $N_N \in R_M$ of the model may give unreliable results in noisy conditions. We propose to further limit the solutions of the two set of parameters in the spaces spanned by polynomial functions.

The further modeling to constraint the values of H_Δ uses the polynomial model. We assume the H_Δ has a value as a function of frequency and the change is not sudden but is smooth and a polynomial of a low order. We assume a noise will follow a Gaussian distribution.

Let $k \in [0, M)$ be the frequency index, and

$$v(k) \underline{\Delta} \frac{k}{M} \in [0,1] \tag{6}$$

be the normalized frequency: In order to reduce the number of free parameters and improve the robustness of the estimation, we further assume that H_Δ is an order-P polynomial function of normalized frequency v:

$$H_\Delta \triangleq \mathrm{diag}\left[b_1^t\theta_H, b_2^t\theta_H, \ldots b_k^t\theta_H, \ldots b_M^t\theta_H\right] \tag{7}$$

where

$$b_k^t \triangleq \left[1, v(k), v^2(k), \ldots v^{P-1}(k)\right] \tag{8}$$

$$\theta_H \triangleq \left[\theta_H^{(1)}, \theta_H^{(2)}, \ldots \theta_H^{P-1}\right]^t \tag{9}$$

Similarly, we assume that N_N is an order-Q polynomial function of normalized frequency.

$$N_N \triangleq \left[c_1^t\theta_N, c_2^t\theta, \ldots, c_k^t\theta_N, \ldots c_M^t\theta_N\right]^t \tag{10}$$

where

$$c_k^t \triangleq \left[1, v(k), v^{(2)}(k), \ldots v^{Q-1}(k)\right] \tag{11}$$

$$\theta_N \triangleq \left[\theta_N^{(t)}, \theta_N^{(2)}, \ldots \theta_N^{(Q-1)}\right]^t \tag{12}$$

The model parameter set is then:

$$\lambda \triangleq \{\theta_H, \theta_N, R\}$$

(13)

Determination of parameters:

When you change H_Δ and N_N in a Gaussian distribution it will change the shape. We change the shape so that the probability of observing Y_N is maximized and is represented by equations 14 and 15.

Polynominal coefficients;

Using maximum likelihood criterion to determine the parameter set λ, we have

$$\frac{\partial}{\partial \theta_H} p(Y_N(1), Y_N(2), \ldots Y_N(T) \mid \lambda)$$

$$= \sum_{i-1}^{T} \sum_{k=1}^{M} \left(Y_N^k(i) - b_k^t \theta_H \, Y_R^k(i) - c_k^t \theta_N \right) \frac{\partial}{\partial \theta_H} \left\{ b_k^t \theta_H \right\} Y_R^k(i) = 0 \qquad (14)$$

$$\frac{\partial}{\partial \theta_N} p(Y_N(1), Y_N(2), \ldots Y_N(T) \mid \lambda)$$

$$\sum_{i=1}^{T} \sum_{k=1}^{M} \left(Y_N^k(i) - b_k^t \theta_H \, Y_R^k(i) - c_k^t \theta_N \right) \frac{\partial}{\partial \theta_N} \left\{ c_k^t \theta_N \right\} = 0 \qquad (15)$$

with

$$\frac{\partial}{\partial \theta_H} \left\{ b_k^t \theta_H \right\} = b_k \qquad (16)$$

$$\frac{\partial}{\partial \theta_N} \left\{ c_k^t \theta_N \right\} = c_k \qquad (17)$$

By interchanging the summations, Equation 14 and Equation 15 can be rewritten as:

$$\forall_p = 1 \ldots P,\ \sum_{j=1}^{P} \theta_H^{(j)} \sum_{k=1}^{M} v^{p+j-2}(k) \sum_{i=1}^{T} Y_R^k(i) + \sum_{j=1}^{Q} \theta_N^{(j)} \sum_{k=1}^{M} v^{p+j-2}(k) \sum_{i=1}^{T} Y_R^k(i)$$
$$= \sum_{k=1}^{M} v^{p-1}(k) \sum_{i=1}^{T} Y_R^k(i) Y_N^k(i) \tag{18}$$

$$\forall_q = 1 \ldots Q,\ \sum_{j=1}^{P} \theta_H^{(j)} \sum_{k=1}^{M} v^{q+j-2}(k) \sum_{i=1}^{T} Y_R^k(i) + T \sum_{j=1}^{Q} \theta_N^{(j)} \sum_{k=1}^{M} v^{q+j-2}(k)$$
$$= \sum_{k=1}^{M} v^{q-1}(k) \sum_{i=1}^{T} Y_N^k(i) \tag{19}$$

Denote:

$$\alpha(m, f, g) \underline{\underline{\Delta}} \sum_{k=1}^{M} v^m(k) \sum_{i=1}^{T} f^k(i) g^k(i) \tag{20}$$

$$\beta(m, f) \underline{\underline{\Delta}} \sum_{k=1}^{M} v^m(k) \sum_{i=1}^{T} f^k(i) \tag{21}$$

$$\gamma(m) \underline{\underline{\Delta}} T \sum_{k=1}^{M} v^m(k) \tag{22}$$

and

$$A_p \underline{\underline{\Delta}} \left| A_p^{(1)}, A_p^{(2)}, \ldots A_p^{(P)} \right| \text{with } A_p^{(j)} = \alpha\left(p + j - 2, Y_R, Y_R\right) \tag{23}$$

$$B_p \underline{\underline{\Delta}} \left| B_p^{(1)}, B_p^{(2)}, \ldots B_p^{(Q)} \right| \text{with } B_p^{(j)} = \beta(p + j - 2, Y_R)$$

(24)

$$C_q \triangleq \left[C_q^{(1)}, C_q^{(2)}, \ldots B_q^{(P)}\right] \text{with } C_q^{(j)} = \beta(q + j - 2, Y_R)$$

(25)

$$D_q \triangleq \left[D_q^{(1)}, D_q^{(2)}, \ldots D_q^{(Q)}\right] \text{with } D_q^{(j)} = \gamma(q + j - 2)$$

(26)

$$\mathbf{u} \triangleq \left[u_1, u_2, \ldots u_p\right]^t \text{with } u_p = \alpha(p - 1, Y_R Y_N)$$

(27)

$$\mathbf{v} \triangleq \left[\upsilon_1, \upsilon_2, \ldots \upsilon_Q\right]^t \text{with } \upsilon_q = \beta(q - 1, Y_N)$$

(28)

Equation 19 and Equation 20 can be expressed as a linear system where H and N parameters are variable as follows:

$$\sum_{j=1}^{P} \theta_H^{(j)} A_p^{(j)} + \sum_{j=1}^{Q} \theta_N^{(j)} B_p^{(j)} = u \ , \ p = 1 \ldots P$$

(29)

$$\sum_{j=1}^{P} \theta_H^{(j)} + \sum_{j=1}^{Q} \theta_N^{(j)} D_q^{(j)} = u_q, \ q = 1 \ldots Q$$

(30)

Or, equivalently:

$$\begin{bmatrix} \mathrm{A} & \mathrm{B} \\ \mathrm{C} & \mathrm{D} \end{bmatrix} \begin{bmatrix} \theta_H \\ \theta_N \end{bmatrix} = \begin{bmatrix} u \\ \mathrm{v} \end{bmatrix}$$

(31)

where

$$A_{(P\times P)} \underline{\underline{\Delta}} [A_1, A_2, \ldots A_P]^t \tag{32}$$

$$B_{(P\times Q)} \underline{\underline{\Delta}} [B_1, B_2, \ldots B_P]^t \tag{33}$$

$$C_{(Q\times P)} \underline{\underline{\Delta}} [C_1, C_2, \ldots C_Q]^t = B^T \tag{34}$$

$$D_{(Q\times Q)} \underline{\underline{\Delta}} [D_1, D_2, \ldots D_Q]^t \tag{35}$$

Equation 31 is a linear system of P + Q variables, and can be solved by a general linear system solution method. See W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C. The Art of scientific programming.* Cambridge University Press, 1988.

Solving linear system equations:

Alternatively, a more efficient solution can be used, which solves one linear equation of order P and another of Q, rather than jointly solving linear systems with (P+Q) variables.

From Equation 31 and Equation 34, we have the following block linear equations:

$$A\theta_H + B\theta_N = \mathrm{v} \tag{36}$$

$$B^t\theta_H + D\theta_N = v \tag{37}$$

From Equation 36,

$$\theta_H = A^{-1}(u - B\theta_N)$$

(38)

From Equation 37 and Equation 38 we obtain a linear system of equation on θ_N:

$$\left(D - B^t A^{-1} B\right)\theta_N = \mathrm{v} - B^t A^{-1} u$$

(39)

Solving Equation 31 can therefore be achieved by first solving Equation 39 for θ_N and then using Equation 38 for θ_H. Similarly, for θ_H, we can derive, from Eq-37,

$$\theta_N = D^{-1}\left(\mathrm{v} - B^t \theta_H\right)$$

(40)

From Eq-36 and Eq-40 we obtain a linear system of equation on θ_H:

$$\left(A - BD^{-1}B^t\right)\theta_H = u - BD^{-1}\mathrm{v}$$

(41)

Solving Equation 31 can therefore also be achieved by first solving Equation 41 for θ_H and then using Equation 40 for θ_N.

Depending on the order of polynomials, one of the two solutions could be computationally more efficient than the other. Finally, we point out the property that A and D are symmetrical and contain only positive elements. That property can be exploited for more efficient solutions.

Covariance matrix:

To solve for the coefficients of the covariance matrix, we make use of the two equalities:

$$\frac{\partial}{\partial A}\log|A| = A^{-1}$$

(42)

$$\frac{\partial}{\partial A} a^t A a = a a^t$$

(43)

To calculate the covariance matrix, we set the derivative of the logarithm of Equation 5 with respect to R-1 to zero, which gives:

$$R = \frac{1}{T} \sum_{i-1}^{T} (Y_N(i) - H_\Delta Y_R(i) N_N)(Y_N(i) - H_\Delta Y_R(i) - N_N)t$$

(44)

To quantify the goodness of the fitting between the model and data, we use:

$$\in = \text{Trace } (R).$$

(45)

Referring to Fig. 2 there is illustrated the process steps for calibration. A voice utterance i is applied to a high quality path and a test path as illustrated in Fig.1 (Step 1) and in Step 2 a measurement is made for each frame. For each frame of each utterance i the power spectrum is determined for reference Y_R and test Y_N After all frames are measured Step 3, for each utterance calculate equations 32-35 and 27 and 28 for A,B,C,D,u and v (Step 4). In Step 5 the noise estimate θ_N and the channel estimate θ_H are calculated using equation 31 or the noise estimate θ_N using equation 39 and channel estimate θ_H is calculated using equation 38.

Referring to the test equipment 10 of Fig.1, the outputs Y_R and Y_N are applied to processor 19 that processes the signals according to Fig. 2 described above to produce the channel estimate θ_H and the noise estimate θ_N signal outputs. These output may be displayed on a display or directly applied to modify acoustic models in a recognizer. For example, the test equipment described in Fig. 1 may be used to test a cellular phone in a car. After a test utterance in the car using the test calibration unit described above the reference signal and the test signal are processed and the equipment provides a channel estimation and noise estimation either on the

display for the user to manually enter the findings in the speech recognizer of the car or the outputs are provided directly to the recognizer.

To show this actually works we generate this kind of environment. The goal is to identify them back. The system has to recover the H and N from the observed Y_N. One is synthesized distortion and the other is recovered.

To simulate a distortion, the test speech signal *yR(n)* is modified in time domain according to the following operation:

$$_{yN}(n) =_{yR} (n) * h_{\Delta}(n) + n_N(n) \tag{46}$$

The speech signal in the example is an utterance of the digit string two *five eight nine six oh four* (5.6 seconds). *ho(n)* is a five-coefficient FIR band-pass filter with cutoff frequencies at 0.10 and 0.35 and *26dB* attenuation. *nN(n)* is a computer generated white Gaussian noise sequence.

In speech signal, the energy is concentrated at low frequencies, which yields higher polynomial fitting error at high frequencies. To balance the errors, speech signal is pre-emphasized. As pre-emphasizing speech signal is a common practice in speech recognition, H_{Δ} and N_N estimated using pre-emphasized speech could be directly usable for compensation.

Throughout the experiments reported,

- we used *9th* order of polynomes for convolutive noises *(P = 9),* and *6th* order of polynomes for additive noises *(Q = 6).*

- Noise estimate shown in the figures (labeled as "noise.PW") are obtained by averaging 30 seconds of noise power spectra.

- In the figures below, "C" stands for convolutional noise, and "A" for additive noise. "POLY" stands for an estimate obtained by polynomial models. "FILTER" stands for the frequency response of the band-pass FIR filter.

To measure the estimation error, we use averaged inner product of the difference between the estimated noise and a reference:

$$e_v \triangleq \frac{1}{M}(\hat{v}-v)'(\hat{v}-v) \qquad \forall_v \in \{H_\Delta, N_N\} \tag{47}$$

- For convolutive noise, the reference $\hat{H}_\Delta$ is the power spectrum of the filter $h_\Delta(n)$.

- For additive noise, the reference $\hat{N}_N$ is the average of power spectra of the noise sequence.

Fig 3 to Fig 6 shows the results of estimation for convolutive and additive noises. In order to test the robustness against additive noise, A white Gaussian noise is introduced to the test channel, with variable SNR.

The results are shown for a wide SNR ranges from 30dB to -6 dB. In each figure, estimate by independentcomponent bias model and by polynomial bias model are shown along with a reference bias. The following can be observed:

- At 30dB, the estimates of convolutive noise by both models are fairly close to the reference (Fig. 3). However, for additive noise, while the polynomial model gives good estimate, the estimate of additive noise by independent-component bias model shows large error with respect to the reference (Fig.4).

To show the relative improvement using polynomial models, Fig.7 and Fig. 8 plot the estimation error according to Equation 47, as function of SNR.

- Fig. 7 shows that the estimation of convolutive bias with independent component model gives 4.7 to 8.4 time larger estimation error than with polynomial models.

- Fig. 8 shows that the estimation of additive bias with independent component model gives 7.5 to 12.3 time larger estimation error than with polynomial models.